Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression

Wimber, Maria; Alink, Arjen; Charest, Ian; Kriegeskorte, Nikolaus; Anderson, Michael C

DOI:
10.1038/nn.3973

Retrieval Induces Adaptive Forgetting of Competing Memories via Cortical Pattern Suppression

Maria Wimber[1,2], Arjen Alink[2], Ian Charest[2], Nikolaus Kriegeskorte[2], Michael C. Anderson[2,3]

[1] *School of Psychology, University of Birmingham, UK*

[2] *MRC Cognition & Brain Sciences Unit, Cambridge, UK*

[3] *Behavioural and Clinical Neurosciences Institute, Cambridge, UK*

Corresponding author:

m.wimber@bham.ac.uk

Maria Wimber

School of Psychology

University of Birmingham

Birmingham

B15 2TT

United Kingdom

Abstract

Remembering a past experience can, surprisingly, cause forgetting. Forgetting arises when other competing traces interfere with retrieval, and inhibitory control mechanisms are engaged to suppress the distraction they cause. This form of forgetting is considered adaptive because it reduces future interference. The impact of this proposed inhibition process on competing memories has, however, never been observed both because behavioural methods are “blind” to retrieval dynamics and because neuroimaging methods have not isolated retrieval of individual memories. Here we introduce a canonical template tracking method to quantify the activation state of individual target memories and competitors during retrieval. This method revealed that repeatedly retrieving target memories suppressed cortical patterns unique to competitors. Pattern suppression was related to engagement of prefrontal regions implicated in resolving retrieval competition, and, critically, predicted later forgetting. We thus demonstrate a cortical pattern suppression mechanism through which remembering adaptively shapes which aspects of our past remain accessible.

Introduction

Remembering, it seems, is a double-edged sword. Research in humans and animals points to the pivotal role that retrieval plays in shaping and stabilizing memories[1,2]. However, the remembering process also induces forgetting of other memories that hinder the retrieval of the memory we seek[1,3,4]. It has been hypothesized that this surprising "dark side" of remembering is caused by an inhibitory control mechanism that suppresses competing memories and causes forgetting; this putative process is adaptive because it limits current and future distraction from competitors[5,6]. Yet, no study has ever directly observed memories as they are suppressed by the hypothesized inhibitory control mechanism. Behavioral methods are, by their nature, blind to the internal processes unfolding during retrieval; and neuroscience has lacked methods capable of isolating neural activity associated with individual memories. In the current fMRI experiment, we tested for the existence of the hypothesized adaptive forgetting process by developing a template-based pattern tracking approach that quantifies the neural activation state of single memory traces. In so doing, we tracked the fate of behaviorally invisible traces, providing a window into the suppression process thought to underlie adaptive forgetting in the human brain.

Our effort to observe the dynamics of adaptive forgetting builds on work examining the neural processes associated with retrieval competition. One approach has used multi-voxel pattern analysis to measure visual cortical activity when a retrieval cue concurrently elicits multiple visual memories. These studies revealed that pattern classifiers have difficulty discriminating whether a retrieval cue is eliciting a memory of a face or an object when both types of content are associated to it, even when only one type of content is to be retrieved[7,8]. It cannot be discerned, however, whether this finding reflects the co-activation of individual memories or of the broad categories to which the memories belong (e.g. faces, objects). A second approach has focused on control mechanisms that resolve retrieval competition by selecting between competing memories. Competition during episodic retrieval engages prefrontal cortical areas associated with selection during semantic retrieval[9]. Specifically, during selective recall of a target memory, ventrolateral prefrontal cortex activity predicts later forgetting of competing memories[5-6,10-11], consistent with the possibility that this area contributes to resolving competition. Together, these two lines of work suggest that lateral prefrontal cortex contributes to adaptive forgetting by exerting a top-down modulatory influence on competing memories in posterior representational areas.

In the present study, we sought to isolate neural indices of individual memory traces, so that we might observe retrieval competition and its resolution as it unfolds in the brain,

and to link these dynamics to adaptive forgetting. To achieve this, we trained participants to associate two images (e.g., Marilyn Monroe and a hat) to each of a set of cue words and then recorded brain activity during a selective retrieval phase in which one of those visual memories (e.g., Marilyn Monroe) was repeatedly retrieved (Fig.1a-b). On each retrieval trial, participants covertly retrieved the first picture they had associated with the cue (henceforth, the target) in as much detail as possible. Across the selective retrieval session, participants retrieved each target four times. Importantly, one quarter of the cue words were set aside, and did not appear in the selective retrieval task. As such, the associations for these cues served as a baseline for assessing the behavioural and neural changes induced by repeated target retrieval.

Our main concern was how retrieving the target affected the competing memory associated with the same cue (henceforth, the competitor). We assumed that the reminder initially would co-activate the target and the competitor, and that resolving this competition in favour of the target would engage inhibitory control to degrade the competitor's neural representation in visual and memory processing regions. We further hypothesized that this degradation would hinder later retrieval of the affected representation, so that on a final visual recognition test, participants should be worse at discriminating inhibited pictures from similar lures, compared to their discrimination accuracy for baseline pictures (Fig.1a).

Our primary goal was to track the suppression of individual memories in visual and memory processing regions. Tracking competitor suppression required a way to discern evidence, during selective retrieval, that the neural pattern associated with a target or its competitor was reactivated. To achieve this, we had participants perform a perceptual localizer task (not shown in Fig. 1a), in which they viewed a subset (50%) of the target, competitor and baseline pictures, multiple times. For each picture, we derived a canonical multivariate activity pattern, representing the perceptual trace that it typically evoked. We assumed that this canonical signature pattern might resemble the visual memory formed during encoding, and provide a template for assessing objectively how much, during each retrieval trial, the visual memory was reactivated. Indeed, previous findings[12–14] indicate that episodic retrieval reinstates perceptual traces established during encoding in late visual processing areas. Memory-unique representations also have been observed in the hippocampus during retrieval[15]. Together, these findings suggest that it may be possible to isolate individual memory patterns in visual and memory processing areas during retrieval, and use them to track the dynamics of selective retrieval.

We therefore hypothesized that across repeated recall trials, as retrieval became more successful and complete, the reactivated pattern in visual and memory processing

regions would become increasingly similar to the canonical template of the target being retrieved. Memory-unique target reactivation during each retrieval trial would be present when the pattern measured on that trial resembled the target template (e.g. Marilyn Monroe) more than it resembled baseline templates from the same category (e.g. Albert Einstein). Critically, if inhibitory control degrades competing memories, the neural pattern during target recall should grow progressively less similar to the canonical template of that target's competitor. Memory-unique competitor suppression during each retrieval trial would be present if similarity of the measured pattern to the specific competitor (e.g. hat) template is driven below its similarity with baseline templates from the same category (e.g. goggles).

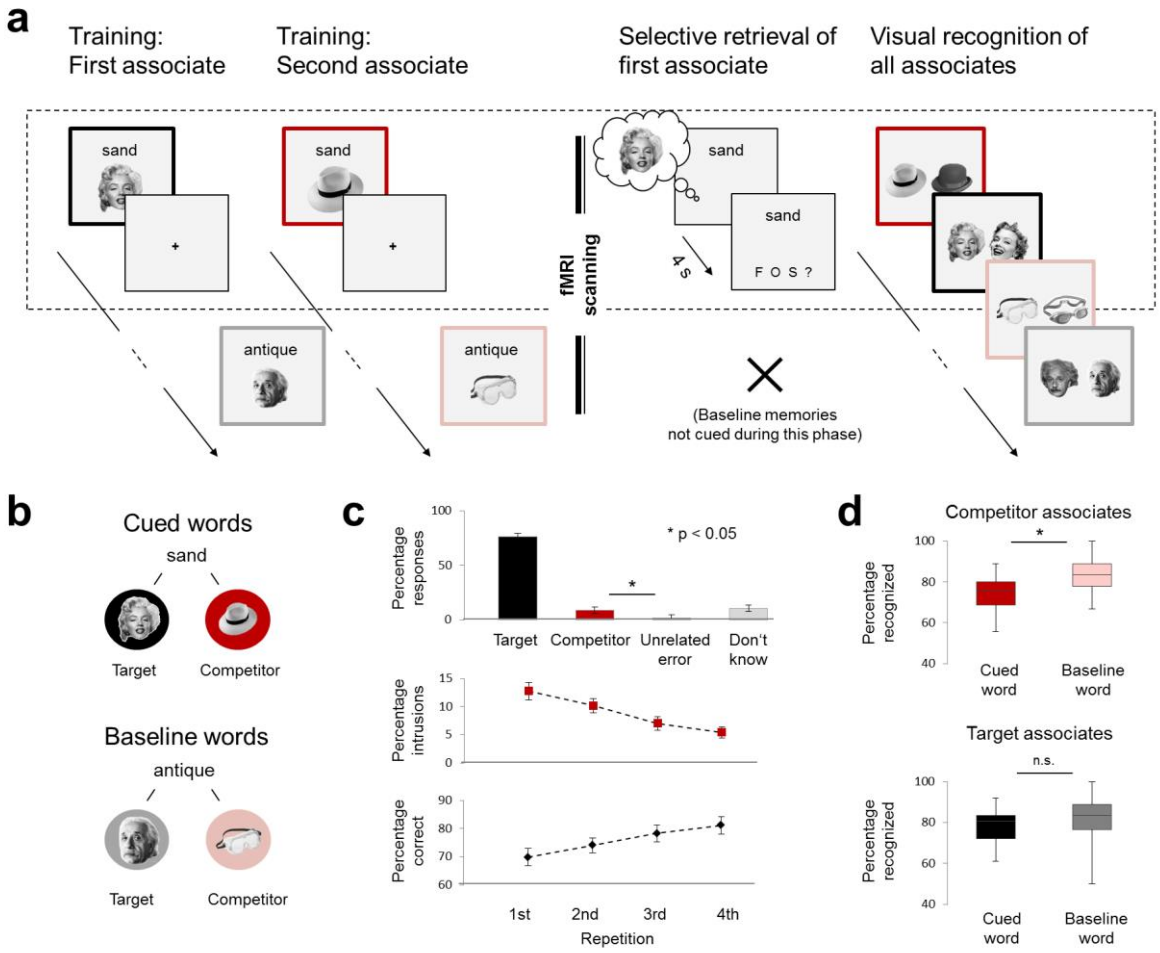

Figure 1. Schematic of the procedure (excluding initial familiarization and the pattern localizer) and behavioural results. *(a) Participants were trained on novel word-picture pairs, each word being linked with two associates. During scanning, participants were cued with a word (4 times each across the entire selective retrieval task), and were asked to retrieve the first associate that they studied (the target), with the second associate (the competitor) assumed to interfere. On each trial they classified the memory that came to mind as being a face ("F"), object ("O"), scene ("S"), or unsuccessful retrieval ("?"). Some of the originally trained targets were not tested during this phase, and served as a baseline against which we assessed the impact of selective target recall. We expected to observe a disruptive aftereffect of selective retrieval on competing associates on a*

forced-choice visual recognition task that required participants to distinguish studied pictures from familiar foils. The coloured frames illustrate item types and were not visible to participants. (b) Illustration of the associative relationships assumed to have been formed after training, and of the different types of items created by the experimental procedure. (c) Behavioural data from the selective retrieval phase. The upper panel shows the proportion of trials on which participants correctly selected the category of the target (e.g. face), or incorrectly selected the category of the competitor (e.g. object), the third (unrelated) category not linked to the current cue word (e.g. scene), or "don't know." The lower panels show the number of intrusion errors (competitor responses) and the corresponding increase in correct responses across repetitions (mean +/- SEM) (d) Behavioural results from the visual recognition memory task. The upper panel shows the disruption of discrimination performance for competitors, compared with their matched baseline items. Boxes reflect median (+/- 1st and 3rd quartile, error bars show minimum and maximum).

Results

Performance during initial training. Training of the first and second associates to each cue occurred in learning-test cycles outside the scanner (Online Methods). During training, first associates were recalled at 77.1% (SEM=2.9%) in the first retrieval cycle, and at 86.4% (SEM=2.6%) in the second. The second associates were recalled at 70.7% (SEM=3.0) in their first and only retrieval cycle.

Performance during selective retrieval. Selective retrieval was performed in the scanner. Because on each trial, participants classified which category of memory they retrieved, we could determine whether they had recalled the correct target category. *P*articipants selected the correct category for the target on 74.7% (SEM=2.9%) of the trials (Fig.1c). When they made errors, they selected the competing picture's category significantly more often (mean=9.2%, SEM=1.1%) than the third, unrelated category (mean=2.3%, SEM=0.3%; t_{23}=6.53, p<0.001). These competitor intrusion errors varied across the four repetitions ($F_{3,69}$=21.8, p<0.001; Fig. 1c), showing a linear decline ($F_{1,23}$=55.4, p<0.001). This pattern is consistent with the possibility that inhibitory control rendered competitors less interfering over repetitions.

Selective retrieval induces forgetting of competitors. As a first step, we tested whether presenting an item's cue during retrieval had different effects on recognition performance depending on whether an item was a first or second associate. A 2 by 2 repeated-measures ANOVA with the factors ITEM TYPE (cued vs. baseline) and ASSOCIATE (first vs. second) revealed a significant interaction ($F_{1,23}$=4.70, p=0.041). Posthoc t-tests confirmed that selective retrieval reduced later recognition of competitors (mean=75.2%, SEM=17.6%) (t_{23}=4.91, p<0.001), compared to recognition of corresponding

baseline items from the second training set (mean=82.1%, SEM=17.1%) (Fig.1d). Thus, remembering the targets induced forgetting of competing memories (irrespective of their category, see Supplementary Fig.1), in line with past work[1,4]. Interestingly, below-baseline forgetting correlated, across individuals, with the number of intrusions observed during selective retrieval (R=0.39, p=0.030), consistent with the idea that retrieval-induced forgetting arises from a control process that reduces interference.

In contrast, recognition of targets (mean=78.6%, SEM=16.7%) did not differ reliably from recognition of corresponding first-studied baseline items (mean=79.7%, SEM=23.9%; t_{23}=0.57, p=0.713), providing little evidence for retrieval-based enhancement. Recognition of the two types of baseline items (first and second associates) did not differ reliably (t_{23}=0.93; p=0.362). Overall, results from the visual recognition test confirmed that selectively recalling target memories disrupts later memory for competitors, supporting the possibility that inhibitory control disrupted competitors' visual-episodic representations.

Imaging Results

Measuring the reactivation of unique memories. In a new canonical pattern tracking approach, we quantified changes in activation of each unique target and competitor across repeated retrievals (see Fig. 2 for rationale). We hypothesized that ventral visual cortex and the hippocampus would carry item-specific information about retrieved content[12–15], and that ventral visual regions would also show strong categorical reactivation[7,8,12]. At the end of scanning, we presented half of the trained pictures 6 times each in a one-back task (see Online Methods for rationale). From this, we constructed canonical multivariate templates based on the average voxel-wise activity pattern elicited by each picture (e.g., Marilyn Monroe). These templates gave us a neural standard against which to assess how much a visual memory was reactivated during selective retrieval.

To quantify item-specific reactivation, we correlated (using Pearson coefficients) the observed neural pattern elicited on each retrieval (e.g. cuing participants with the word "sand" in the examples in Fig.1 and 2) with the current target template (e.g. Marilyn Monroe), and with the current competitor template (e.g. the hat). Importantly, we also computed templates for baseline pictures (e.g. Albert Einstein, and goggles). These baseline templates allowed us to quantify how much the specific neural patterns representing the target (e.g. Marilyn Monroe) and the competitor (e.g. hat) were reinstated during a retrieval trial, above and beyond categorically matched baseline items. All selective retrieval trials for which item-

specific templates were available were analysed (Supplementary. Fig.2 reports the same results excluding incorrect retrievals; exact statistics available on request).

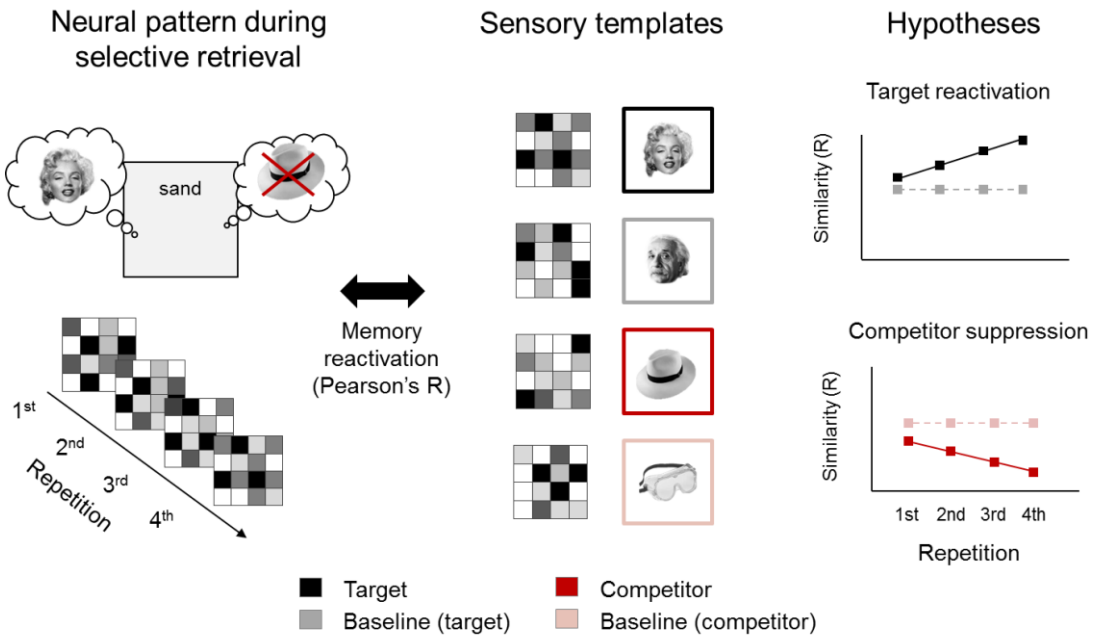

Figure 2. Rationale of the item-specific canonical pattern analysis approach. *For each region of interest, we extracted multivoxel activity patterns elicited during a given selective retrieval trial (left), and computed similarity with the canonical neural templates obtained from the sensory pattern localizer (middle). Item-specific similarity was assessed by correlating the selective retrieval pattern in a given region of interest, trial-by-trial, with the item-unique template of the current target, the template of the current competitor, and the templates of baseline items that were initially trained and came from the same categories as the target and competitor, respectively, but were never cued by a reminder word during the selective retrieval phase. The graphs show the hypothesized changes in pattern similarity across the four repeated retrieval trials. As sketched in these graphs, we expected the patterns during target retrieval to show increasing similarity with the target template (e.g. Marilyn Monroe), compared with baseline first associates from the same category (e.g. Albert Einstein), and decreasing similarity with the competing template (e.g. hat), relative to baseline second associates from the same category (e.g. goggles).*

Emergence of item-unique target patterns. Both ventral visual cortex and the hippocampus showed evidence for target-unique memory reinstatement (Fig.3). Specifically, similarity of the observed pattern with the target template, relative to same-category baseline templates, showed a significant (positive) linear trend across repetitions in both regions of interest (ventral visual cortex: $F_{1,23}$ =12.97, p=0.002; hippocampus: $F_{1,23}$ =11.91, p=0.002; Fig.3), as tested in a repeated-measures ANOVA with the factors ITEM TYPE (target vs.

baseline) and REPETITION (one to four). There was a significant ITEM TYPE by REPETITION interaction in ventral visual cortex ($F_{3,69}$=4.15, p=0.009) and hippocampus ($F_{3,69}$=4.72, p=0.007). Post-hoc tests showed that on the final (fourth) recall attempt, target reactivation exceeded baseline in the hippocampus (t_{23}=2.50, p=0.010), whereas ventral visual cortex showed significant target reactivation on the third (t_{23}=2.01, p=0.028) but not on the fourth repetition (t_{23}=1.44, p=0.082; Fig.3). Neural patterns during retrieval therefore suggest that the unique memory was reinstated increasingly over repetitions, one of the few demonstrations that a memory-specific cortical trace can be elicited by an associatively linked cue (see[14,16] for related findings).

Suppression of unique neural patterns representing competing memories. Next, we correlated the observed pattern during each selective retrieval trial to the competitor's template. Strikingly, across the four repetitions, memory-specific competitor activation showed a significant (negative) linear trend in ventral visual cortex ($F_{1,23}$=10.52, p=0.004) but not in hippocampus ($F_{1,23}$=1.07, p=0.312; note that the hippocampus showed a trend towards suppression when including correct trials only, see Supplementary Fig.2). The ITEM TYPE by REPETITION ANOVA revealed a significant interaction in ventral visual cortex ($F_{3,69}$=3.71, p=0.016) but not the hippocampus ($F_{3,69}$=0.52, p=0.670). Thus, unlike target reactivation, competitor activation in ventral visual areas declines significantly across repeated retrievals.

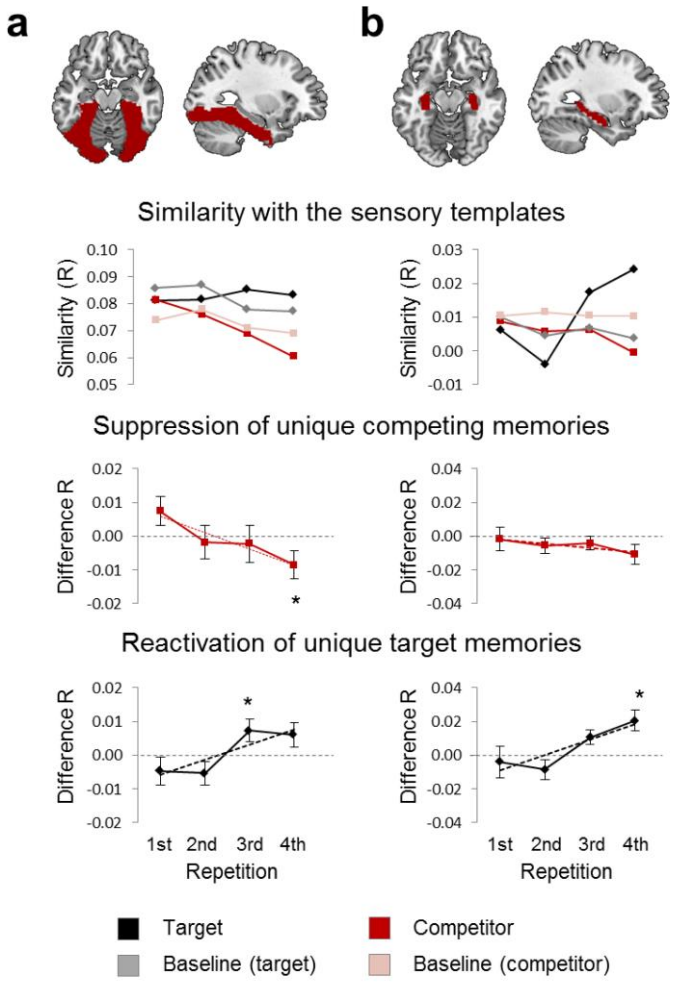

Figure 3. Item-specific target reactivation and competitor suppression. *The multivoxel pattern during selective retrieval was extracted and compared with the sensory template patterns in (a) ventral visual cortex, and (b) hippocampus. The first row shows an overlay of the respective anatomical ROIs on a standard MNI brain. The second row shows the raw average correlation (similarity) between selective retrieval activity and the canonical template of the current target (black solid), the templates of non-cued baseline items from the target category (grey dotted), the current competitor (red), and the templates of non-cued baseline items from the competitor category (pink dotted). Along the x-axis, changes in similarity across the four repetitions of retrieving the same target memory are shown. The third row shows mean competitor-related (red) similarity, subtracting similarity with the respective baseline templates (mean +/- SEM across single subject estimates), along with the average of the best linear fit (ML estimates) across participants. The bottom row shows the same baseline-corrected measures for target-related (black) similarity. Evidence for item-specific memory reactivation or suppression is indicated by a significant ($p < .05$, indicated by asterisks) deviation from zero difference.*

We considered the possibility that this negative trend simply reflects target reactivation becoming more successful and complete, such that the cue would grow more likely over repetitions to selectively elicit the target. If so, competitor reactivation would decline across trials, but cease at a baseline level where the probability of the cue eliciting

the competitor would match its probability of eliciting baseline memories. Conversely, if inhibition suppresses interfering memories during retrieval, similarity between the selective retrieval pattern and the competitor template should decrease significantly below the level of non-cued baseline memories. Supporting the latter, the difference between competitor and baseline similarity (Fig.3, middle)showed a trend towards competitor re-activation during the first retrieval in ventral visual cortex (t_{23}=1.70, p=0.050), but not in the hippocampus (t_{23}=0.13, p=0.449), irrespective of whether we excluded incorrect trials (Supplementary Fig.2). By the final (fourth) repetition, however, similarity with the competitor's template was driven below similarity with same-category baseline templates in both regions (ventral visual cortex: t_{23}=2.14, p=0.022; hippocampus: t_{23}=1.97, p=0.030). These findings indicate that reminders initially activate competitors, but competitors are progressively suppressed below baseline, consistent with the hypothesized inhibition process.

Competitor suppression predicts adaptive forgetting. If inhibition disrupts competing traces during retrieval, our index of cortical competitor suppression should predict adaptive forgetting. Confirming our hypothesis, the extent to which participants down-regulated the competing neural patterns in ventral visual cortex across repetitions predicted below-baseline forgetting of competing memories on our recognition test, (R=–.35, p=0.047; Fig.4). No significant correlation was observed in the hippocampus (R=0.17, p=0.217).

We also tested whether pattern suppression predicted which individual memories would be forgotten. To do this, we derived, for each participant, a measure of pattern suppression for every individual competitor by fitting a linear regression to the decrease in its similarity to its template across the four retrieval trials, relative to baseline similarity (Fig.3, dotted lines). These fits yielded maximum-likelihood (ML) estimates of the slope of the best fitting regression line for each competitor that quantifies its pattern suppression. Consistent with the linear trend analysis, below zero estimates were found in ventral visual cortex (t_{23}=3.33, p=0.001) but no the hippocampus (t_{23}=1.03, p=0.157). We then tested if these memory-specific estimates predicted whether items were forgotten, using logistic regression. In ventral visual cortex, items showing more pattern suppression were indeed more likely to be forgotten (β=5.38, p=0.037). Together, these findings support the hypothesis that cortical pattern suppression underlies adaptive forgetting.

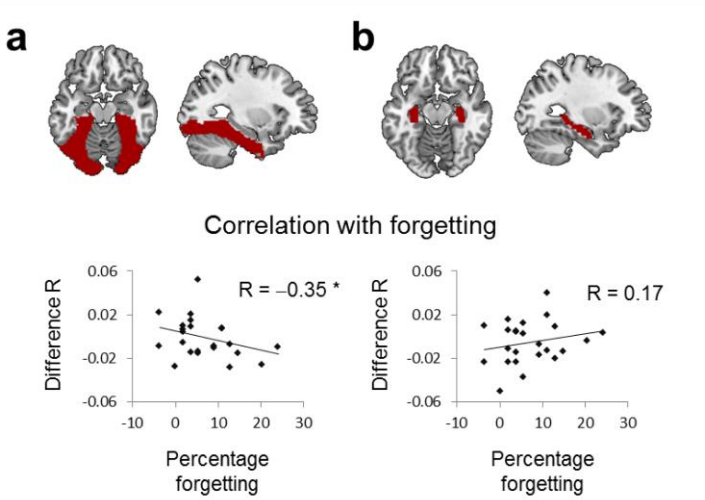

Figure 4. Correlation between item-specific competitor suppression and forgetting. *Across-participant correlations between cortical and behavioural suppression of competing memories are shown separately for (a) ventral visual cortex and (b) the hippocampus. The x-axis in each graph shows our behavioural forgetting index on the delayed visual recognition memory test (forgetting of competitors relative to baseline items, with positive scores indicating more forgetting), and the y-axis shows the overall cortical suppression of competitors during the selective recall task, calculated as the difference between reactivation of competitors and baseline items, averaged across all four repetitions.*

The role of prefrontal cortex in cortical pattern suppression. The prefrontal cortex is a key candidate region for the source of the top-down control signal that induces pattern suppression[5,10,11]. To test this possibility, we defined prefrontal regions of interest based on a functional comparison between early and late selective retrieval trials[5]. The rationale behind this contrast is that demands on the control mechanism should decrease across repetitions as interference is reduced. Replicating past work on retrieval-induced forgetting[5,10,11], this contrast revealed clusters in left and right mid-ventrolateral prefrontal cortex and the inferior frontal junction (including middle and inferior frontal gyri, Fig. 5a; left BA 6/8: xyz = -48 5 43, k = 635 voxels, t_{peak} = 5.73; right BA 9: xyz = 48 11 31, k = 332 voxels, t_{peak} = 5.42).

To test for a role of prefrontal cortex in pattern suppression, we first correlated participants' prefrontal activity during selective retrieval with their slope of competitor suppression (average ML estimate). Critically, average beta estimates in both prefrontal regions-of-interest strongly predicted the slope of competitor suppression in visual cortex (left IFG: R=–0.65, p<0.001; right IFG: R=–0.48, p=0.009; Fig.5b). No relationship was found between prefrontal activity and the slope of target up-regulation (left IFG: R=0.25, p=0.124; right IFG: R=–0.10, p=0.324). The correlation of prefrontal activity with competitor suppression was more negative than its correlation with target enhancement in left IFG

(Hotelling's t_{21}=4.58, p<0.001), and marginally so in right IFG (Hotelling's t_{21}=1.52, p=0.072). Critically, we also tested whether the prefrontal activity during the selective retrieval of individual memories predicted pattern suppression (ML estimate) for that memory's competitor, within participants. Higher prefrontal cortex activity was indeed related to greater pattern suppression (left IFG: R=–0.123; p=0.008; right IFG: R=–0.104; p=0.021).

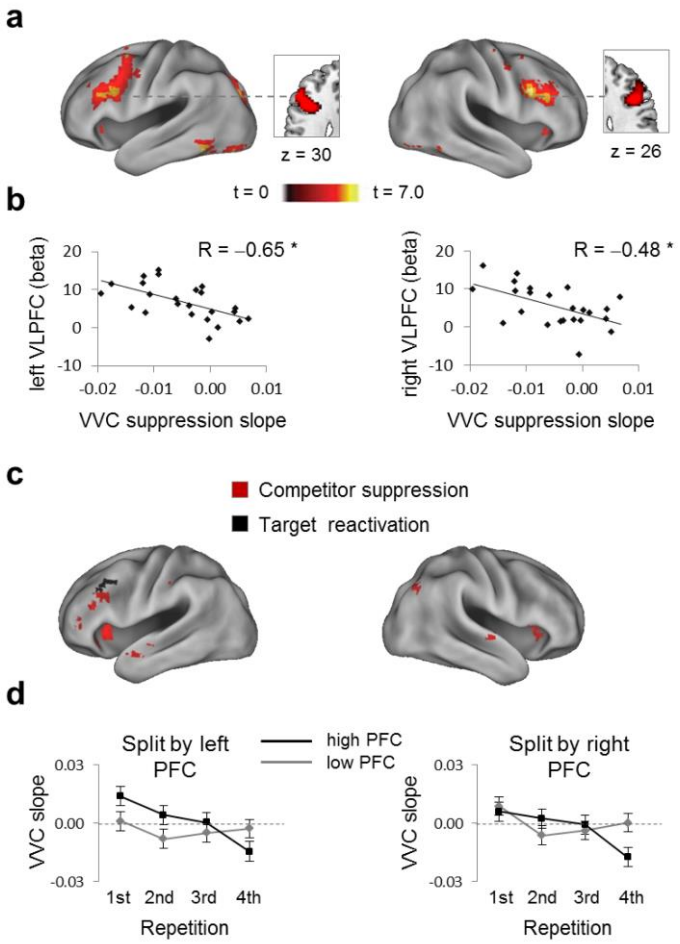

Figure 5. Relationship between prefrontal activity and cortical suppression of competing memories. *(a) Left and right mid-ventrolateral prefrontal cortices (VLPFC) showed stronger univariate activity (p < .001) during early (first half) than during late (second half) selective recall repetitions. (b) The univariate decrease across repetitions in both regions predicted the slope of cortical pattern suppression (ML estimates) in ventral visual cortex (VVC), with larger prefrontal decreases associated with more negative-going slopes of competitor suppression. (c) Whole-brain regression showing areas that, across participants, significantly correlate with the slope of competitor suppression (red) and the slope of target reactivation (black) in ventral visual cortex. Both contrasts are shown at p < .001 (uncorrected). (d) Cortical pattern suppression as a function of PFC engagement, splitting the sample into participants with high and low PFC engagement. Participants with high PFC engagement showed a significant (p < .05) difference in the slope of competitor suppression, and in the level of competitor suppression on the fourth (final) retrieval trial. (e) Error bars in panels (b) and (d) represent SEM across participants for each single measure.*

To further illustrate the link between prefrontal recruitment and pattern suppression, we median split our sample based on prefrontal recruitment (Fig.5d). Participants with high right PFC engagement showed steeper suppression slopes (t_{22}=1.77, p=0.045), and more competitor suppression on the fourth retrieval (t_{22}=2.31, p=0.015). This split revealed no difference in the slope of target enhancement (t_{22}=0.29, p=0.387), nor target reactivation on the fourth retrieval (t_{22}=1.41, p=0.086). Similar patterns were observed when splitting the sample by left PFC. These analyses support a specific functional relationship between PFC recruitment and competitor suppression in visual cortex.

Finally, a whole brain analysis identified several clusters that predicted pattern suppression (Fig.5c, red), mostly in left and right prefrontal cortices (Supplementary Table2). Only one small cluster in the left middle frontal gyrus predicted target enhancement (Fig.5c, black). Together, our results support the possibility that the mid-VLPFC is a source of top-down inhibitory modulation that suppresses the cortical patterns of competing memories.

Voxels diagnostic of competitor activation are suppressed. The evidence for cortical pattern suppression thus far could arise because of at least two factors: because competitor patterns become noisier, or because inhibition truly suppresses diagnostic features of the competitor (i.e., the “hat” voxels). We hypothesized that the latter would be the case[17], and sought to isolate voxels diagnostic of a given target or competitor. We first used item-specific linear pattern classifiers to isolate voxels that most reliably distinguished individual targets or competitors from their respective control items during the sensory pattern localizer. In a second step, we computed changes in signal strength of only the 10% of voxels in our ventral visual cortex mask that were most diagnostic for each target and competitor, as determined by linear weights of the trained classifiers (Online Methods; Supplementary Fig.3 to see how the findings change with voxel diagnosticity).

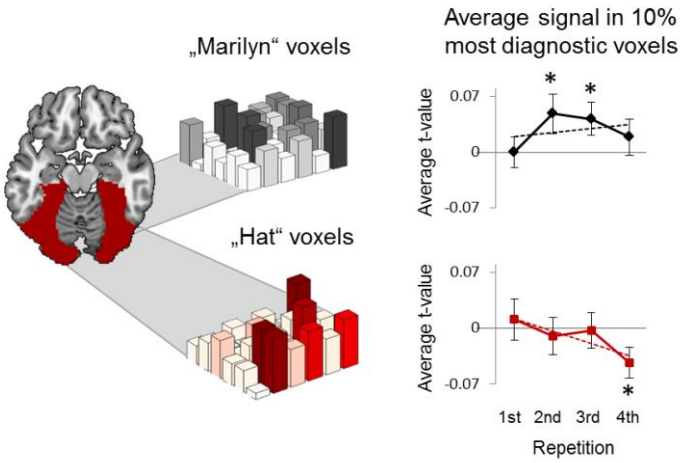

Figure 6. *Activation in diagnostic voxels for individual targets and competitors, across repetitions. Diagnostic voxels were determined from item-specific linear classifiers that were trained to distinguish a given target and a given competitor picture from all same-category baseline items. Based on the weights of these classifiers, we investigated average BOLD signal changes in the 10% most diagnostic voxels of each target (black) and competitor (red). Diagnostic target voxels showed above-baseline activation on the second and third repetitions (upper right). Importantly, on average, competitor voxels showed a significant linear decrease in activation across the four recall repetitions, and a significant below-baseline suppression effect at the final repetition (lower right).*

Having identified diagnostic voxels for each target and competitor, we extracted average activation (t-values) and tested whether activity in those voxels was enhanced for targets, and importantly, suppressed for competitors (Fig.6). Unexpectedly, target voxel activity showed no positive linear trend across repetitions ($F_{1,23}$=0.47, p=0.500), and no significant above-baseline activation on the final repetition (t_{23}=0.80, p=0.216). However, consistent with our inhibition hypothesis, voxels diagnostic of the competitor showed a significant linear decrease across repetitions ($F_{1,23}$=5.48, p=0.028) and significant below-baseline suppression (t_{23}=2.10, p=0.023). A significantly negative competitor slope was only obtained in the 10% most diagnostic voxels (Supplementary Fig 3). These findings suggest that cortical pattern suppression is at least partly driven by reduced activity in voxels that contribute strongly to representing competing memories.

Categorical target reactivation without competitor suppression. To underscore the advantages of our item-unique analyses, we conducted two categorical analyses that assessed whether patterns during selective retrieval showed reactivation of the target or competitor categories. For the similarity analysis (Fig.7), we calculated a template for each category (e.g., a face template) based on baseline pictures from the localizer. Categorical similarity was assessed by computing the correlation between the pattern observed during each retrieval trial and the template of that trial's target category, its competing category, and its non-involved (categorical baseline) category.

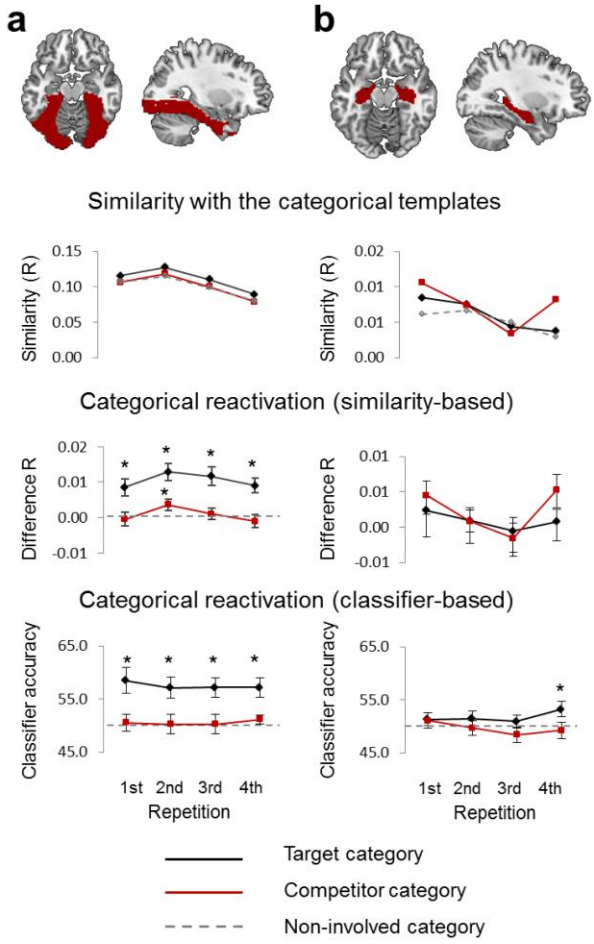

Figure 7. *Results from the categorical multivariate analyses in (a) ventral visual cortex and (b) the hippocampus. The upper line plots show raw similarity (Pearson correlation) values between selective recall patterns and the canonical template of the target category (black solid), the canonical template of the competing category (red solid), and the canonical template of the currently non-involved category (grey dotted), averaged across trials and participants. The middle plots show the same measures transformed into differences in categorical activation relative to category that was not involved on a given trial. The lower row shows the results from a complementary categorical analysis using linear pattern classifiers (SVMs), with plotted means reflecting classifier accuracy in determining the target and competitor category (against the baseline, non-involved category). Both approaches converge in indicating highly significant categorical target reactivation in ventral visual cortex (but not the hippocampus), with no reliable change over repetitions. No significant below baseline suppression of the competitor's category was evident. All measures plotted as mean +/- SEM (across subjects).*

Ventral visual cortex but not hippocampus showed strong evidence for categorical target activation (main effect target vs. baseline in ventral visual cortex: $F_{1,23}$=29.79, p<0.001; hippocampus: $F_{1,23}$=0.96, p=0.338) that did not reliably change with repetition (interaction with repetition in ventral visual cortex: $F_{3,69}$=1.60, p=0.196; hippocampus: $F_{3,69}$=0.43, p=0.732; Fig.7). We observed similar results with a categorical analysis based on linear machine learning algorithms (Fig.7, bottom; Online Methods): Classification of the target

category across recall trials was above chance in ventral visual cortex (t_{23}=4.88, p<0.001) and the hippocampus (t_{23}=2.38, p=0.013), and showed stable categorical reactivation across repetitions, with no linear trend (ventral visual cortex: $F_{1,23}$=0.11, p=0.750; hippocampus: $F_{1,23}$=0.65, p=0.428). This high above-chance similarity/classification mirrors classification responses collected during the selective retrieval phase, which were accurate from the first repetition (Supplementary Fig.4 illustrates how classifier performance varies depending on the response given). Importantly, participants' classification responses during selective retrieval are only diagnostic as to the accuracy of the category retrieved, not the specific item, like the classifier output itself.

Critically, despite strong target activation, categorical patterns did not detect competitor suppression. Activation of competitor categories did not significantly differ from baseline in either region of interest (main effect of competitor vs. baseline in ventral visual cortex: $F_{1,23}$=0.63, p=0.437; hippocampus: $F_{1,23}$=3.80, p=0.064), and showed no interaction with repetition (ventral visual cortex: $F_{3,69}$=1.43, p=0.240; hippocampus: $F_{3,69}$=2.50, p=0.067). The linear classifier analysis confirmed this pattern, showing a trend toward above-chance classification of the competitor category when averaged across repetitions in ventral visual cortex (t_{23}=2.00, $p_{two\text{-}tailed}$=0.057) but not in the hippocampus (t_{23}=0.59, $p_{two\text{-}tailed}$=0.561). Classification performance showed no linear decrease across repetitions (ventral visual cortex: $F_{1,23}$=0.21, p=0.651; hippocampus: $F_{1,23}$=1.00, p=0.328). Finally no relationships were found between activation of competitor categories (overall, or slope over repetitions, between- or within participants) and forgetting (all p>0.15, exact statistics available on request). These results suggest that the inhibitory mechanism underlying adaptive forgetting suppresses features of individual competing memories, not global categorical patterns.

Discussion

Remembering does not merely reawaken memories of the past; it has a "darker side" that induces forgetting of other experiences that interfere with retrieval, dynamically altering which aspects of our past remain accessible. Remembering, quite simply, causes forgetting. It has been hypothesized that this adaptive forgetting process is caused by an inhibitory control mechanism that suppresses distraction from competing memories[1,3–5]. Five key findings indicate that we have, for the first time, isolated the hypothesized adaptive forgetting mechanism and shown it to be implemented by the suppression of distributed neocortical patterns that represent competing memories.

First, selective retrieval caused forgetting of competing memories. When we repeatedly cued participants to retrieve target items, competing memories were recognized less well later on, compared to baseline items (Fig.1d). This effect occurred for images of faces, objects, or scenes, indicating a domain-general process. Forgetting was observed on a forced-choice recognition test that displayed the putatively inhibited visual item, reducing memory search demands. Observing below-baseline forgetting even though our test provided potent, vivid, item-unique cues indicates that retrieval disrupts the sensory features of competing memories[17,18]—a possibility compatible with an adaptive forgetting process that suppresses visual cortical patterns underlying those memories. Critically, forgetting was predicted by the tendency of competitors to interfere, as reflected by how often participants mistakenly selected the competitor's category during selective retrieval trials. This tendency of competitors to intrude reduced gradually over retrieval trials (Fig. 1c), consistent with an active suppression process. Taken together, these findings exhibit the hallmarks indicating a role of inhibitory control in retrieval-induced forgetting, supporting the possibility that we succeeded in eliciting the putative adaptive forgetting process.

Second, during the four selective retrievals, cortical pattern indices revealed that competing memories were measurably reactivated and then progressively suppressed (Fig.3). Our reactivation index measures how much the activation pattern elicited by the cue resembled the perceptual template for the associated target or competitor memories, and provides an objective neural standard for quantifying the retrieval of individual memories. Gradual suppression of competing patterns is expected based on the hypothesized inhibitory control mechanism thought to underlie adaptive forgetting.

It was essential to consider whether the decline in competitor activation over target retrievals might reflect processes other than cortical pattern suppression. For example, participants may grow efficient at reinstating the target over repeated retrievals, reducing the chances of reactivating competitors. Alternatively, an associative unlearning mechanism, in which target retrievals punish competing associations, may make the cue less likely to reactivate competitors[1]. Both alternatives predict, however, that the competitor's activation should simply approach the level observed for baseline memories, and never decline below baseline because, even if cue-competitor associations were unlearned entirely (or, alternatively, if the cue became perfectly efficient at eliciting the target) the cue should merely fail to reactivate the competitor; it should be as if the competitor is unassociated to the cue, like baseline items. Inhibition, by contrast, predicts that competitors are actively inhibited, and that their cortical traces will be suppressed below the activity observed for baseline items.

This prediction was confirmed. This third key finding—below baseline pattern suppression—provides encouraging and distinctive support for the hypothesized inhibition mechanism.

Even if inhibition caused pattern suppression, this finding does not establish the relevance of these reductions to adaptive forgetting. Two final findings support an active forgetting interpretation, and establish important characteristics of cortical pattern suppression. First, if inhibitory control reduced mnemonic activation by acting on cortical sites representing competitors, this putative footprint of inhibition should be predicted by activation in prefrontal regions implicated in inhibitory control. Such a finding would distinguish an adaptive mechanism that acts during goal-directed retrieval from other, incidental mechanisms that may weaken memories. For example, reactivating memories briefly during tasks unrelated to retrieval[16,19,20], may strengthen or weaken the "reawakened" memories depending on how active they become. This forgetting is predicted by a computational model of inhibition[21], and is proposed to not require control by the prefrontal cortex. In contrast, we found that the engagement of mid-ventrolateral prefrontal regions previously linked to adaptive forgetting[5,10,11] predicted pattern suppression in ventral visual cortex both across and within participants, with more robust VLPFC engagement predicting greater pattern suppression (Fig. 5b-d). These findings support a contribution of mid-VLPFC to a top-down control signal that suppresses competition in visual cortex.

Fifth, if reduced competitor activation in ventral visual cortex is relevant to adaptive forgetting, it should predict forgetting. This relationship was observed: participants showing the steepest decline in competitor activation showed the most forgetting (Fig.4); and even within participants, those individual memories showing the steepest suppression slope were most likely to be forgotten. These relationships support the possibility that cortical pattern suppression plays an instrumental role in adaptive forgetting.

Taken together, these five key findings provide strong and specific support for the hypothesized cortical pattern suppression process, and for its role in producing adaptive forgetting in the human brain. Our findings suggest further properties of pattern suppression that may prove important if corroborated. For instance, our canonical pattern tracking approach allowed us to investigate how inhibition modulates cortical traces. Does inhibition target the unique cortical pattern causing interference (i.e., the "hat pattern"), or the global representation of the competing category (i.e., an "object pattern")? Several findings favour an item-specific suppression mechanism. First, pattern suppression for individual items was driven, in part, by down-regulated activity in voxels distinguishing a competitor from other members of its category and from the target (Fig.6). These findings are expected based on models of memory inhibition[17], according to which inhibition targets features representing a

competitor that do not overlap with those representing the target. Second, despite robust categorical reactivation of targets, the competitor's category showed no evidence of suppression. In line with previous studies[7,8], categorical patterns even showed a trend in the opposite direction, with early retrievals showing co-activation of the competitor's and target's categories. Thus, although categorical activations can reveal competition, our results indicate that the brain's adaptive response to resolving competition– inhibition–suppresses a competitor's diagnostic features distinguishing it from other exemplars of its category, and from the memory being retrieved.

A second interesting observation is that hippocampal patterns exhibited weaker evidence for pattern suppression, despite robust target reactivation. Weaker competitor suppression may be relevant to computational models of hippocampal-neocortical processing, assuming that the hippocampus, in contrast to neocortex, uses sparse coding and efficiently separates overlapping patterns[22,23]. If the neocortical components of a distributed memory are more disrupted by competition[23,24], it may be functional for inhibitory control to target neocortical areas to suppress interference. These speculations about the selectivity of pattern suppression to neocortex must remain tentative, awaiting further confirmation.

The proposed top-down mechanism that supports selective retrieval by suppressing competing memories parallels mechanisms believed to support visual selective attention and visual working memory[25–30]. Selective attention enhances targets and suppresses distracting information, a pattern demonstrated from single neurons up to EEG and BOLD activity[31–34], and such adaptive modulations of sensory regions are believed to be driven by lateral prefrontal cortex[34,35]. Recent studies indicated a causal role of the inferior frontal junction in exerting this top-down influence[34]. This frontal area overlaps with regions implicated in resolving mnemonic competition in previous work[5,10,11] and, critically, in the present study. By showing a relationship between prefrontal activity and competitor suppression, our findings reinforce theoretical parallels between the mechanisms the brain uses to resolve mnemonic competition on the one hand, and sensory competition on the other hand[28], building a theoretical bridge spanning attention and long-term memory.

Studying the neural basis of forgetting has proven challenging because the substrate of episodic memories (the "engram") has been difficult to pinpoint in brain activity. By capitalizing on the relation between perception and memory, the present study detected neural activity sensitive to the activation of individual memories. This canonical pattern tracking approach provided a unique window into the invisible neurocognitive processes triggered when a reminder recapitulates several competing memories in neocortex.

Strikingly, we could track dynamic changes in the activity of individual memories during selective retrieval, as competition was resolved. In doing so, we established clear evidence for cortical pattern suppression as a key mechanism of adaptive forgetting in the human brain. More broadly, this work converges with a growing literature showing that forgetting often serves an adaptive function[2,36]; it establishes how, by simply using our memory system via selective retrieval, we adapt the landscape of memory to the demands of mental life.

Author Contributions

M.W. and M.C.A. designed the experiment, with important contributions from I.C. and N.K. M.W. conducted the experiment. M.W., A.A. and I.A. analysed the data. All authors contributed to the analysis approach and to data interpretation. M.W. and M.C.A. wrote the manuscript.

Acknowledgments

We thank Bernhard Staresina and Simon Hanslmayr for commenting on previous versions of the manuscript. This work was supported by a fellowship from the German Research Foundation (WI3784/1-1) awarded to M.W., and by UK Medical Research Council grant MC-A060-5PR00 awarded to M.C.A.

Online Methods

Participants. Twenty-four healthy participants (20 female) aged 20 – 32 years (mean 24.2 years) were recruited from the MRC CBU volunteer panel. They all had normal or corrected-to-normal vision, and reported no history of neurological or psychiatric disease. The experimental procedure was conducted in accordance with the local ethics review board, including the requirement of written informed consent from each participant before the beginning of the experiment.

Materials. The word material used as verbal cues consisted of 72 English words drawn from the MRC linguistic database (http://www.psych.rl.ac.uk/). Words were selected on the basis of having relatively low imageabilty (mean = 571.3, SD = 37.3) and concreteness (mean = 545.1, SD = 54.6) ratings such that they would not elicit concrete mental images by themselves when presented to participants in the scanner. Pictures were 144 photographs of well-known faces, well-known scenes, and everyday objects (48 pictures per category) from a range of in-house databases as well as the internet. All images were converted to black-and-white and scaled to cover the same visual angle. Note, however, that faces and objects were background stripped and thus contained extensive areas of white background, while scenes always covered the full angle of the picture. In addition to the materials used in the main experimental runs, three additional words and six additional pictures were used for demonstration purposes during practice runs outside the scanner. The 144 pictures were split into two sets of 72 pictures each (24 per category). One set was trained together with a cue word as first associates, and the other set was trained together with the same cue words as second associates. The two associates linked to the same cue word always came from different categories (e.g. a face and an object, see Fig.1). Fifty-four pictures out of the 72 first associates (18 per category) later became the to-be-retrieved targets, and 54 pictures out of the 72 second associates later became competitors. The remaining 36 pictures (18 first associates, 18 second associates) were linked to cue words that never appeared during the scanned selective retrieval task and thus served as baselines for the targets and competitors, respectively. Assignment of pictures to conditions was counterbalanced such that across participants, each picture equally often served as a target, competitor and baseline item.

Experimental procedure. Familiarization with the pictures, and the training on word-picture associations was carried out in a separate testing room outside the scanner. The first task was a familiarization phase, during which participants were presented with all 144 pictures

used in the experiment as well as their corresponding similar lures (used in the visual recognition test, see below), and thus saw a set of 288 pictures in random order. Each picture appeared alone first; followed by its verbal label (e.g. “Charlie Chaplin”) after 1 s, the label remaining on the screen for another 1.5 s. Participants indicated with a button press whether they recognized (i.e., were familiar with) the face, object or scene shown on the photograph. In cases in which they indicated that they were unfamiliar with an item, the same picture was presented to them for a second time at the end of the familiarization phase.

After familiarization, participants were trained on the first set of 72 word-picture associations. To facilitate learning, the training was separated into 3 blocks, each consisting of an initial learning, a test, and a re-test cycle for 24 out of the 72 word-picture pairs. At the beginning of each block, participants were presented with the 24 word-picture pairs for 4.5 s each (4 s + 0.5 s inter-stimulus interval). The word was shown above the picture, and it was emphasized to participants that they should make an effort to memorize the picture in as much detail as possible in order to be able to bring back a vivid mental image of the picture when cued with the word, later in the scanner. In order to build strong links between the words and the pictures, we instructed participants to use a mental imagery strategy, that is, to use the word and picture in an interactive way (e.g., use the cue word to make the picture move, change colour etc.). This initial learning was followed by two cycles of test-feedback practice. On each trial, participants first saw a word (e.g. “sand”) on a blank screen, and were asked to orally provide the label (or a short description) of the picture they had learned to associate with this word. Two similar versions of the correct picture associate (the same versions also used in the later visual recognition test) appeared 3 s later, and participants had to indicate which of the two pictures they had previously linked with the word. This procedure was again aimed at emphasizing the encoding of as many visual details as possible.

After finishing training on the first set of pictures (which would become the targets during later selective retrieval), participants were instructed that they would now be trained on a second set of associates for each word (which would become the competitors during selective retrieval), and that later in the scanner they might need either of the two associates. It was emphasized to participants that they would be required to retrieve the two associates separately, and should thus not inter-relate the two pictures associated with the same cue word (i.e. they should not form an integrated mental image). We did so because integration between competing memories has been shown to be a main factor limiting retrieval-induced forgetting[1]. In terms of the procedure, training of the second set of associates (which would later become the competitors) was performed exactly as for the first set, with the exception

that the test-feedback practice involved only one instead of two cycles. After training of the second set, participants were given a short practice on the tasks they would perform in the scanner.

During the recall task in the MRI scanner, participants were prompted with a cue word for 4 s each, followed by a response prompt ("F – O – S - ?") asking them to indicate the category of the picture they were currently recalling (fingers 2-5 of the right hand corresponding to "face", "object", "scene", and "don't know", respectively). The response prompt was presented for 1.5 s (ISI = 1 s). Feedback was given as soon as participants pressed a button, with the correct response option lightening up in green colour. We instructed participants to always press a button while the response prompt was still present on the screen, because they would miss the feedback when responding too late. However, responses given during the following inter-stimulus interval were still included in the data analysis. The selective recall task was followed by a short (~ 2min) period of rest, followed by the final recognition test. In this task, each trial presented participants with two similar pictures, both of which had been presented before in the familiarization phase, but only one of which they had initially been linked with a cue word. Importantly, the cue words were not shown during the final test. The two pictures were presented simultaneously, to the left and right of the fixation cross, for 3.5 s (ISI = 1 s). Participants used their right index and middle finger to select the picture they had linked with a word during training.

The final task conducted in the scanner was a pattern localizer for individual pictures, conducted to obtain the item-unique sensory templates. During the localizer, the BOLD activity pattern in response to a subset of 72 of the initially trained 144 pictures was sampled (only half of the items were sampled due to time constraints). The subsample of pictures was chosen randomly for each participant, with the constraint that it had to include 18 target pictures, the 18 corresponding alternative associates from the same word-picture triples, 18 baseline pictures that had been trained as first associates, but were not recalled during the selective recall task, and the 18 corresponding alternative associates from the same word-pictures triples. The latter two picture types were used to obtain baseline templates to compare the targets and competitors, respectively, against. Each of the sampled pictures was presented 6 times overall. Picture presentation occurred in the context of a one-back task, where each picture was shown for 1.5 s (ISI = 1 s) and participants were instructed to respond with their index finger as fast as possible whenever two consecutive items in the picture sequence were the same.

The sensory templates were sampled at the end of the scanning phase for several reasons. First, the localizer overall lasted for ~25 min, and we did not want to introduce a

delay of this length between study of the word-picture pairs and the selective retrieval task. Second, and more importantly, one might expect a priori that the similarity between the recall patterns and the sensory templates would become higher with increasing temporal proximity between the localizer and the time at which the templates are sampled. Such an increase could occur simply because any neural pattern sampled at a given time during scanning would show a drift towards or away from the localizer patterns depending on how far in time from the localizer it is sampled. Based on such pattern drifts, recall patterns should overall become less similar to the sensory templates if the localizer is conducted before the selective recall phase; and more similar to the templates if the localizer is conducted at the end of the experiment, after selective recall. Because our main effect of interest in this study was an effect of *decreasing* similarity across retrieval repetitions (for the competitors), it was a more conservative approach to conduct the localizer at the end of the experiment, such as to not risk the effect to be confounded with spurious similarity decreases caused by pattern drifts. Note that such spurious similarity changes might, according to this reasoning, have affected the increasing similarity we found with the sensory templates for target representations. Having said this, we believe that it is unlikely for all our effects to be caused by spurious correlation through pattern drifts, because of the use of very well controlled baseline measures. In particular, pattern drifts towards the "template state" should have affected the similarity with all templates, including the sensory templates of control items.

However, one might still argue that differences inherent in the localizer templates may affect the overall correlation between the neural patterns during selective retrieval and the different types of templates. We took several measures to minimize this concern, the results of which are shown in Supplementary Table 2 and Supplementary Fig. 5. These analyses showed that the templates did not significantly differ in signal-to-noise ratio (SNR; computed as mean t-value across all voxels in the template divided by the standard deviation), in informational content as measured by Shannon entropy, or in the degree to which they correlated with other templates from the same condition ("correlationability"). Importantly, because the aim of these analyses was to show *no difference* between conditions (i.e. between target templates and their respective baseline templates, and between competitor templates and their respective baseline templates), Supplementary Table 1 also reports Bayes factors[37] together with the p-values, giving an indication of the strength of evidence in favour of the null hypothesis.

For all tasks conducted in the scanner, event sequences were optimized for rapid event-related designs using self-programmed MATLAB code, based on the genetic algorithm suggested by Wager and Nichols[38]. For the multivoxel pattern localizer, the output of the

algorithm was modified to obtain a reasonably high number of picture repetitions (11-15% of the trials), as to keep participants engaged in the one-back task. In each of the scanned tasks (selective retrieval, visual recognition, and the pattern localizer), events were interspersed with null-trials (fixation periods covering the same period as actual events) corresponding to one-third of the overall trial number.

fMRI data acquisition and pre-processing. Imaging data were acquired on a 3 tesla Siemens Trio scanner using a 32-channel head coil. High-resolution (1 mm^3 isotropic voxels), T1-weighted anatomical scans were acquired at the beginning of each session using a magnetization-prepared rapid acquisition gradient echo (MP-RAGE) sequence resulting in 192 sagittal slices. Functional volumes were obtained in three separate sessions corresponding to the recall phase (772 volumes), the final picture discrimination test (274 volumes), and the picture localizer (727 volumes). Functional volumes consisted of 32 axial slices (3.75 mm slice thickness, 3 x 3 mm in-plane resolution) covering the full brain, and were acquired using a descending T2*-weighted echo-planar imaging (EPI) pulse sequence (repetition time = 2.0 s; echo time = 30 ms, flip angle = 78°). The first 5 volumes of each session were discarded to allow for stable tissue magnetization.

SPM8 (www.fil.ion.ucl.ac.uk/spm/) was used for pre-processing and univariate analyses. For all analyses, images were slice timed and realigned in space to the first image of each session, and global effects within each session and voxel were removed using linear detrending[39]. All multivariate analyses were conducted in native (subject) space without normalizing or smoothing the EPI images.

Univariate data analysis. For univariate analyses, EPI images were additionally normalized (using the segmentation algorithm as implemented in SPM8) and smoothed with an 8mm full-width-at half-maximum (FWHM) Gaussian kernel. Events of interest were modelled as delta (stick) functions and convolved with a first-order canonical hemodynamic response function (HRF). Button presses were included in all single-subject models as events of no interest, and the movement parameters from spatial realignment were included as nuisance variables. For univariate group statistics, single-subject activation maps of each condition of interest were entered into a within-subject ANOVA using pooled errors. The main comparison of interest between early and late retrieval trials (Fig. 5a) was calculated within this ANOVA, and results are reported on an uncorrected p-level of <.001 (minimum extent threshold k = 10 voxels). For the regression analysis reported in Fig. 5c, an activation map contrasting early and late retrieval trials was calculated in each single participants, and entered into a whole-brain, group-level GLM using multivariate indices of target enhancement and competitor suppression (see below) as linear regressors.

Similarity-based multivariate data analysis. A template-based variant of representational similarity analysis (RSA[40,41]) was used to assess the degree to which the neural patterns that were active during recall were similar to the neural pattern templates obtained from the pattern localizer. To this end, each trial and repetition during selective retrieval was modelled as a single event (regressor) in a general linear model by convolving a delta stick function at the onset of the event with a canonical HRF. For obtaining the sensory templates, the six repetitions of the same item as visually presented during the pattern localizer were modelled as one event (regressor). For the item-specific linear pattern classification analysis (Fig.6), we modelled the six repetitions of each item as separate regressors. With respect to selective retrieval activity, each retrieval trial was modelled as a single event (regressor). Overall, this procedure produced 54 (items) x 4 (repetitions) t-maps from the selective retrieval task, and 72 t-maps from the pattern localizer. Only the 18 x 4 recall patterns for which item-specific localizer templates were available were included in the item-specific analysis, whereas all 54 x 4 recall patterns were included in the categorical analysis.

Anatomical regions of interests (ROIs) were built based on the human atlas as implemented in the WFU pickatlas software (http://fmri.wfubmc.edu/software/PickAtlas), and back-projected into native space using the inverse normalization parameters obtained from SPM during segmentation. The large ventral visual cortex ROI was comprised of bilateral inferior occipital lobe, parahippocampal gyrus, fusiform gyrus, and lingual gyrus (all bilateral and based on AAL definitions). The hippocampal ROI contained only the bilateral hippocampi, based on the Talairach Demon's brodmann areas (dilated by a factor of 2 as this yielded optimal coverage of our individual subjects' anatomies). The multivariate patterns used in the correlation approach were obtained by extracting the raw beta values from each region of interest and in response to each event of interest, converting them to t-values and finally vectorizing these t-values[42,43]. All similarity-based analyses were based on a correlation approach, using Pearson correlation as a metric of similarity between the sensory canonical templates and selective retrieval activity.

For the item-specific RSA analysis, we computed the correlation between each single selective retrieval trial and the corresponding target template (yielding an index of target reactivation), and the correlation between the same trial and the corresponding competitor template (yielding an index of competitor reactivation). To obtain an appropriate baseline for target and competitor reactivation on each single trial, we computed the correlation between the selective retrieval pattern and each single baseline template corresponding to the same category as the target (used as a baseline for item-unique target reactivation), or the same

category as the competitor (used as a baseline for item-unique competitor reactivation). For the target and competitor baseline measures, correlations were first computed between the retrieval pattern and each single available baseline template from the target's and competitor's category, respectively. We then used the average correlation with the baseline templates (as opposed to the correlation with the average baseline template, which is an important difference) as a measure of baseline similarity. All further analyses performed on the raw similarity values, including linear fits, are described in the results section of the main text.

For the categorical analysis, we first computed an average face template, an average object template, and an average scene template based on all available baseline pictures from the pattern localizer task. To assess categorical target enhancement and competitor suppression, we then correlated each selective retrieval trial with the categorical template of the current target category (e.g. a face), the categorical template of the current competitor category (e.g. an object), and the average template of the category that was currently not involved as target or competitor category (e.g. a scene). All methods using linear pattern classifiers are described in the Supplementary Methods.

Repeated measures ANOVAs and t-tests were used to test for differences in multivariate pattern similarity. All t-tests were used to test directional hypotheses, and unless indicated otherwise, one-tailed t- and p-values at an alpha threshold of 0.05 are thus consistently reported throughout the results section. Brain-brain and brain-behaviour relationships were tested both within- and across subjects. For across-subjects relationships, Spearman correlation coefficients were used. All within-subject, item-by-item correlations (including logistic regression) were computed from fixed-effects models in order to increase power to detect a relationship. Note that for the correlations between prefrontal cortex and neural suppression slopes, the same results were obtained when using a random-effects model; for the logistic regression relating neural suppression slopes to behavioural outcome on the final test, there were not enough forgotten trials on an individual subject basis to yield stable beta coefficient estimates. For reasons of consistency, were therefore report fixed-effect analyses throughout. Before collapsing trials across subjects, outlier trials were identified within each subject, and rejected according to an absolute deviation from the mean (with a criterion of 2.0[44]).

Classifier-based multivariate analyses. All pattern classification analysis used linear support vector machines as implemented in the LIBSVM library (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). For the "diagnostic voxels" analysis reported in the main results section and in Figure 6, we trained separate binary classifiers, based on the six

repetitions of each item during the sensory pattern localizer, to distinguish an individual target and competitor item from each same-category baseline item. For example, to derive the linear weights that optimally separate the "hat" pattern in ventral visual cortex from the pattern elicited by other baseline objects, six binary classifiers were trained to distinguish the hat from the goggles, the hat from a chair etc. During this procedure, each voxel is assigned a linear weight (ω), the absolute value of which directly reflects the importance of a feature (voxel) in discriminating the two classes. We defined the intersection of those voxels that consistently yielded the 10% highest weights across the separate classifiers for each competitor/target as the "diagnostic" voxels for a given target or competitor. The same procedure was used to determine the diagnostic voxels for each baseline item, except that here we trained five binary classifiers for each item, separating this baseline item from all remaining, same-category baseline items.

Having derived these diagnostic voxels for each localizer item, we were then able to compute the average activity (average t-values) of the voxels most diagnostic for the target and competitor item or a given recall trial during the selective retrieval task. In order to ensure that the diagnostic target and competitor voxels did not overlap, we also removed the intersection of those two sets of voxels for this analysis. This rationale was purely theory-driven, as competitor voxels ("features") that overlap with target voxels (features) should not be subject to inhibition[1]. Finally, to parallel the similarity-based analyses using our template tracking approach, we subtracted from those average activity estimates on each retrieval trial the average activity of other voxels that are diagnostic for same-category baseline items, but not the specific target and competitor items involved in this trial. The results of this analysis are described in the main results section, and depicted in Figure 6.

For our categorical classification analysis, we trained binary linear classifiers purely on the patterns elicited in ventral visual cortex by the baseline items during the sensory pattern localizer. Three separate classifiers were trained to optimally distinguish faces from scenes, faces from objects, and objects from scenes. We then tested the accuracy of those classifiers to guess, on each selective retrieval trial, the category of the target by using the binary classifier representing the target vs. non-involved, baseline category (e.g. the face-scene classifier for the examples shown in Figures 1 and 2), and to guess the category of the competitor by using the binary classifier representing the competitor vs. non-involved, baseline category (e.g., the object vs. scene classifier in the example shown in Figures 1 and 2). Note that this way of setting up the analysis automatically builds in the non-involved category, i.e. the one category that should not be elicited by a given cue word, as a baseline on each trial. The results reported in the main results section and in Figure 7 (lower row) and

Supplementary Figure 5 correspond to the average accuracy, across all 54 retrieval trials, to predict the target (black) and competitor (red) categories, respectively.

References

1. Anderson, M. Rethinking interference theory: Executive control and the mechanisms of forgetting. *J Mem Language* **49**, 415–445 (2003).
2. Hardt, O., Einarsson, E.O. & Nader, K. A bridge over troubled water: reconsolidation as a link between cognitive and neuroscientific memory research traditions. *Annu Rev Psychol* **61**, 141–167 (2010).
3. Anderson, M.C., Bjork, R.A. & Bjork, E.L. Remembering can cause forgetting: retrieval dynamics in long-term memory. *J Exp Psychol Learn Mem Cogn* **20**, 1063–1087 (1994).
4. Storm, B.C. & Levy, B.J. A progress report on the inhibitory account of retrieval-induced forgetting. *Mem Cognit* **40**, 827–843 (2012).
5. Kuhl, B.A., Dudukovic, N.M., Kahn, I. & Wagner, A.D. Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nat Neurosci* **10**, 908–914 (2007).
6. Wimber, M. *et al.* Neural markers of inhibition in human memory retrieval. *J Neurosci* **28**, 13419–13427 (2008).
7. Kuhl, B.A., Bainbridge, W.A. & Chun, M.M. Neural reactivation reveals mechanisms for updating memory. *J Neurosci* **32**, 3453–3461 (2012).
8. Kuhl, B.A., Rissman, J., Chun, M.M. & Wagner, A.D. Fidelity of neural reactivation reveals competition between memories. *Proc Natl Acad Sci USA* **108**, 5903–5908 (2011).
9. Badre, D. & Wagner, A.D. Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia* **45**, 2883–2901 (2007).
10. Wimber, M., Rutschmann, R.M., Greenlee, M.W. & Bäuml, K.-H. Retrieval from episodic memory: neural mechanisms of interference resolution. *J Cogn Neurosci* **21**, 538–549 (2009).
11. Wimber, M. *et al.* Prefrontal dopamine and the dynamic control of human long-term memory. *Transl Psychiatry* **1**, e15 (2011).
12. Polyn, S.M., Natu, V.S., Cohen, J.D. & Norman, K.A. Category-specific cortical activity precedes retrieval during memory search. *Science* **310**, 1963–1966 (2005).
13. Ritchey, M., Wing, E.A., LaBar, K.S. & Cabeza, R. Neural similarity between encoding and retrieval is related to memory via hippocampal interactions. *Cereb Cortex* **23**, 2818–2828 (2013).
14. Staresina, B.P., Henson, R.N.A., Kriegeskorte, N. & Alink, A. Episodic reinstatement in the medial temporal lobe. *J Neurosci* **32**, 18150–18156 (2012).
15. Chadwick, M.J., Hassabis, D., Weiskopf, N. & Maguire, E.A. Decoding individual episodic memory traces in the human hippocampus. *Curr Biol* **20**, 544–547 (2010).
16. Poppenk, J. & Norman, K.A. Briefly cuing memories leads to suppression of their neural representations. *J Neurosci* **34**, 8010–8020 (2014).

17. Anderson, M.C., Green, C. & McCulloch, K.C. Similarity and inhibition in long-term memory: evidence for a two-factor theory. *J Exp Psychol Learn Mem Cogn* **26**, 1141–1159 (2000).
18. Spitzer, B. Finding retrieval-induced forgetting in recognition tests: a case for baseline memory strength. *Front Psychol* **5**, 1102 (2014).
19. Kim, G., Lewis-Peacock, J.A., Norman, K.A. & Turk-Browne, N.B. Pruning of memories by context-based prediction error. *Proc Natl Acad Sci USA* **111**, 8997–9002 (2014).
20. Detre, G.J., Natarajan, A., Gershman, S.J. & Norman, K.A. Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia* **51**, 2371–2388 (2013).
21. Norman, K.A., Newman, E., Detre, G. & Polyn, S. How inhibitory oscillations can train neural networks and punish competitors. *Neural Comput* **18**, 1577–1610 (2006).
22. Alvarez, P. & Squire, L.R. Memory consolidation and the medial temporal lobe: a simple network model. *Proc Natl Acad Sci USA* **91**, 7041–7045 (1994).
23. Norman, K.A. & O'Reilly, R.C. Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychol Rev* **110**, 611–646 (2003).
24. Hardt, O., Nader, K. & Nadel, L. Decay happens: the role of active forgetting in memory. *Trends Cogn Sci* **17**, 111–120 (2013).
25. Chun, M.M. & Johnson, M.K. Memory: enduring traces of perceptual and reflective attention. *Neuron* **72**, 520–535 (2011).
26. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu Rev Neurosci* **18**, 193–222 (1995).
27. Gazzaley, A. & Nobre, A.C. Top-down modulation: bridging selective attention and working memory. *Trends Cogn Sci* **16**, 129–135 (2012).
28. Anderson, M.C. & Spellman, B.A. On the status of inhibitory mechanisms in cognition: memory retrieval as a model case. *Psychol Rev* **102**, 68–100 (1995).
29. Kastner, S. & Ungerleider, L.G. The neural basis of biased competition in human visual cortex. *Neuropsychologia* **39**, 1263–1276 (2001).
30. Suzuki, M. & Gottlieb, J. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nat Neurosci* **16**, 98–104 (2013).
31. Gazzaley, A., Cooney, J.W., McEvoy, K., Knight, R.T. & D'Esposito, M. Top-down enhancement and suppression of the magnitude and speed of neural activity. *J Cogn Neurosci* **17**, 507–517 (2005).
32. Cerf, M. *et al.* On-line, voluntary control of human temporal lobe neurons. *Nature* **467**, 1104–1108 (2010).
33. Seidl, K.N., Peelen, M.V. & Kastner, S. Neural evidence for distracter suppression during visual search in real-world scenes. *J Neurosci* **32**, 11812–11819 (2012).
34. Zanto, T.P., Rubens, M.T., Thangavel, A. & Gazzaley, A. Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nat Neurosci* **14**, 656–661 (2011).
35. Squire, R.F., Noudoost, B., Schafer, R.J. & Moore, T. Prefrontal contributions to visual selective attention. *Annu Rev Neurosci* **36**, 451–466 (2013).
36. Nader, K. & Hardt, O. A single standard for memory: the case for reconsolidation. *Nat Rev Neurosci* **10**, 224–234 (2009).

37. Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev* **16**, 225–237 (2009).

38. Wager, T.D. & Nichols, T.E. Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *Neuroimage* **18**, 293–309 (2003).

39. Macey, P.M., Macey, K.E., Kumar, R. & Harper, R.M. A method for removal of global effects from fMRI time series. *Neuroimage* **22**, 360–366 (2004).

40. Kriegeskorte, N. Pattern-information analysis: from stimulus decoding to computational-model testing. *Neuroimage* **56**, 411–421 (2011).

41. Kriegeskorte, N. & Kievit, R.A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn Sci* **17**, 401–412 (2013).

42. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci* **2**, 4 (2008).

43. Misaki, M., Kim, Y., Bandettini, P.A. & Kriegeskorte, N. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage* **53**, 103–118 (2010).

44. Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J Exp Soc Psychol* **49**, 764–766 (2013).